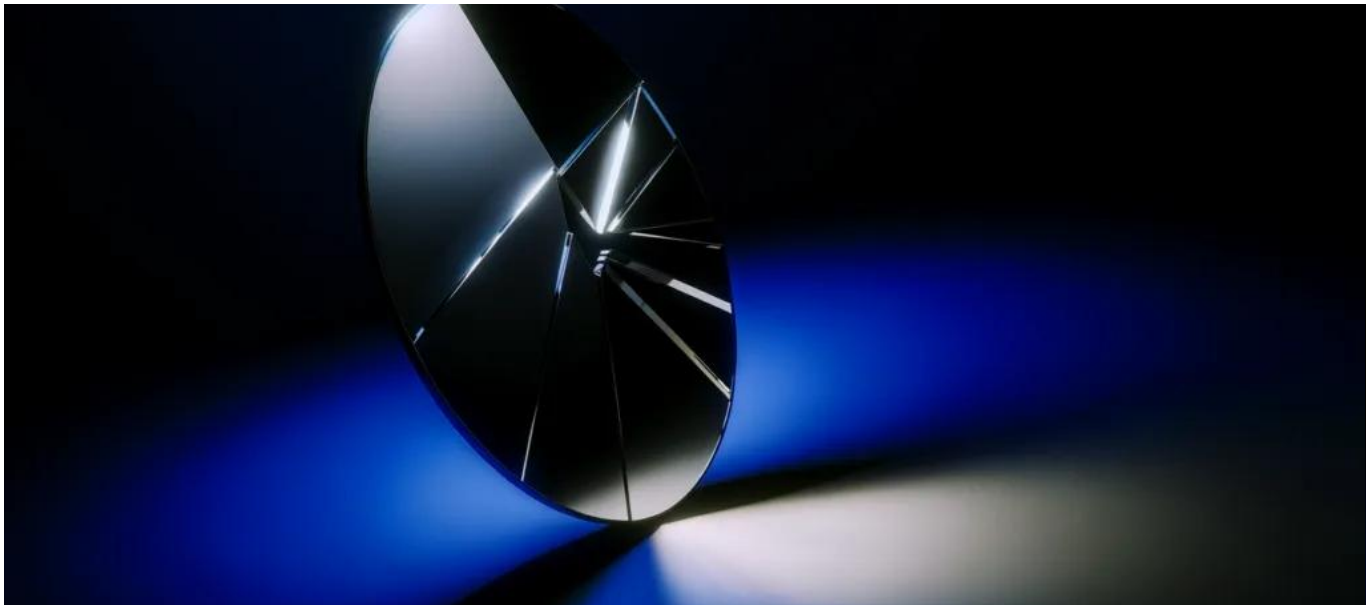


Snorkel AI Commits $3M to Open Benchmarks Grant—Targeting the ‘Biggest Blind Spot’ Where AI Models Excel on Tests But Fail in Production

Claude Opus 4.6 just scored 76% on MRCR v2—up from 18.5% on its predecessor. GPT-5.3-Codex hit 77.3% on Terminal-Bench 2.0. Neither score tells you whether these models will actually work in your production environment.

The News: $3M to Fix How We Measure AI

Snorkel AI announced a $3 million Open Benchmarks Grants program between February 11-13, 2026, with one explicit goal: closing the evaluation gap where agentic AI development has outpaced our ability to reliably measure it. The timing wasn’t accidental—the announcement came less than a week after both Anthropic and OpenAI dropped flagship releases on February 5, each claiming state-of-the-art benchmark scores.

The program doesn't write checks. Instead, it provides in-kind services: compute credits, data access, and research collaboration through a coalition that includes Hugging Face, Prime Intellect, Together AI, Factory HQ, Harbor, and PyTorch. Applications open March 1, 2026, with quarterly selection cycles.

Here's the catch that makes this interesting: all selected teams must release their work under MIT, Apache 2.0, CC BY 4.0, or CC0 licenses. Snorkel isn't building proprietary evaluation infrastructure. They're funding the creation of public goods that the entire industry can use—and critique.

The program targets six dimensions that current benchmarks completely miss: long-horizon tasks, world-modeling, non-stationary environments, multi-artifact outputs, robustness testing, and what they call "human uplift"—measuring whether AI actually makes humans more effective rather than just completing tasks in isolation.

Why Current Benchmarks Are Broken

The AI industry has a measurement problem that everyone acknowledges and no one has fixed. We've optimized for what's easy to measure rather than what matters.

Consider what happens when you test a coding assistant on a benchmark like Terminal-Bench. The model gets a well-specified problem, a clean environment, and a clear success criterion. Score: 77.3%. Impressive. Now drop that same model into a real codebase with legacy dependencies, ambiguous requirements from a product manager, and the need to coordinate with three other services. Performance craters—not by 10%, but often by 50% or more.

This isn't a secret. Engineers who deploy these systems talk about it constantly. But the gap between benchmark performance and production reliability has become so normalized that we've stopped treating it as a crisis. We've accepted that the numbers on the leaderboard are decorative rather than predictive.

> "Benchmaxxing"—the practice of optimizing specifically for public benchmarks—has become the dominant strategy for AI labs competing for

enterprise contracts and developer attention.

The incentives are obvious. A 5-point improvement on a popular benchmark generates headlines, Twitter threads, and sales calls. A 5-point improvement on "ability to handle ambiguous multi-step tasks in messy real-world environments" generates… nothing, because we don't have a reliable way to measure it.

Snorkel's program documentation identifies five specific failure modes that current evaluation frameworks miss entirely:

- **Long-horizon coherence:** Can a model maintain consistent reasoning across dozens of steps, not just three or four?
- **World-model accuracy:** When the model builds an internal representation of the problem space, how often is that representation wrong in ways that compound?
- **Non-stationary adaptation:** Real environments change. Files get modified. APIs return different responses. How does the model handle state that shifts mid-task?
- **Multi-artifact coordination:** Production tasks often require generating code, documentation, tests, and configuration simultaneously. How well does the model maintain coherence across outputs?
- **Trustworthy uncertainty:** When the model doesn't know something, does it admit it—or does it confidently hallucinate?

None of these dimensions appear in MRCR v2 or Terminal-Bench 2.0. They're hard to operationalize, expensive to evaluate, and don't produce clean single-number scores that fit in a press release.

The Technical Reality: Why This Is Hard

Building better benchmarks for agentic AI isn't just a funding problem. It's a fundamental measurement challenge that touches on some of the thorniest issues in software evaluation.

The Reproducibility Problem

Traditional benchmarks work because they're deterministic. Given the same input, you should get the same output, and scoring that output should be unambiguous. Agentic AI breaks this model completely.

When an AI agent interacts with a real environment—browsing the web, executing code, modifying files—the environment changes. The next run of the benchmark starts from a different state. External services return different results. Network latency varies. These aren't bugs; they're features of the real world that we need our agents to handle.

But they make reproducible evaluation nearly impossible. If Claude Opus 4.6 scores 68% on your task today and 72% tomorrow, is that variance or improvement? You can't tell without running hundreds of trials, which gets expensive fast.

The Specification Problem

Production tasks are underspecified by nature. "Build me a dashboard that shows our key metrics" is a perfectly reasonable request that a good software engineer handles every day. It's also impossible to evaluate objectively without human judgment about what "key metrics" means and whether the resulting dashboard actually serves the user's needs.

Current benchmarks avoid this problem by over-specifying everything. The model knows exactly what success looks like because the benchmark tells it explicitly. This makes measurement clean but destroys ecological validity.

Any benchmark that captures real-world performance needs to include ambiguity—and then needs some mechanism to evaluate whether the model handled that ambiguity reasonably. Human evaluation is the obvious answer, but it's expensive, slow, and introduces its own reliability problems.

The Horizon Problem

The difference between a 3-step task and a 30-step task isn't linear. It's closer to exponential.

In a short-horizon task, small errors are recoverable. The model makes a minor mistake in step 2, notices it in step 3, corrects it. No harm done. In a long-horizon task, those same small errors compound. A slightly wrong assumption in step 4 shapes the approach to steps 5-15, and by step 20, the model is confidently executing a plan that's fundamentally misaligned with the actual goal.

Measuring this requires benchmarks that are deliberately long and deliberately messy—exactly the opposite of what makes evaluation convenient. The MRCR v2 benchmark that Claude Opus 4.6 aced? Most tasks complete in under 10 steps. That's not where models fail in production.

The Ground Truth Problem

For many real-world tasks, there is no single correct answer. A model might generate code that works but violates your team's style conventions. It might produce documentation that's technically accurate but misses the level of detail your users need. It might solve the stated problem while ignoring an unstated constraint that any experienced engineer would have caught.

Binary pass/fail evaluation misses all of this nuance. But nuanced evaluation requires judgment, and judgment doesn't scale.

What Most Coverage Gets Wrong

The predictable narrative around Snorkel's announcement focuses on "AI companies finally taking evaluation seriously." This framing misses the more interesting story.

This Isn't About Better Tests—It's About Changing Incentives

The AI industry doesn't lack smart people who understand evaluation. It lacks incentive structures that reward honest measurement over flattering measurement.

Snorkel's bet is that open-source benchmarks, once established, create accountability that proprietary benchmarks can't. When everyone can inspect the evaluation framework, run it themselves, and identify its limitations, benchmark gaming becomes harder and more

embarrassing when exposed.

This is fundamentally a governance intervention disguised as a technical program. The $3M isn't buying better measurement methodology—it's buying legitimacy for a new set of standards that models will be judged against.

The Partner List Tells the Real Story

Look at who's providing resources: Hugging Face, Prime Intellect, Together AI, Factory HQ, Harbor, PyTorch. These aren't evaluation specialists. They're infrastructure providers who have direct exposure to the gap between benchmark claims and production reality.

When you're running inference infrastructure, you see exactly how often models fail on tasks they supposedly excel at. When you're hosting thousands of models, you notice patterns in what breaks. These companies aren't participating out of altruism—they're participating because unreliable benchmarks create support costs, churn, and customer frustration that hit their bottom line.

The Benchmarks That Matter Won't Be Public

Here's the contrarian take: the most valuable evaluation frameworks that emerge from this program won't be the public benchmarks themselves. They'll be the methodologies that companies adapt internally to measure what matters for their specific use cases.

A financial services company deploying AI for document analysis doesn't need a general-purpose long-horizon benchmark. They need an evaluation framework for their specific document types, their specific accuracy requirements, and their specific failure modes. The open-source outputs from Snorkel's program will serve as templates and starting points, not final solutions.

The real value is in the pattern—not the specific benchmark, but the approach to building benchmarks that capture what production performance actually requires.

Practical Implications: What You Should Actually Do

If you're deploying AI systems in production, here's how to think about evaluation in light of this development.

Stop Trusting Headline Benchmarks

This should be obvious, but it bears repeating. When a vendor tells you their model scores X% on benchmark Y, your response should be: "What does that predict about performance on my specific tasks?"

Usually, the honest answer is: "Very little." Benchmark performance correlates weakly with production performance for most enterprise use cases. The correlation is weakest for exactly the tasks where AI is most valuable—ambiguous, multi-step, real-world problems.

Build Your Own Evaluation Sets

The most reliable predictor of how a model will perform on your tasks is how it performs on your tasks. This sounds circular, but it's actionable.

Start collecting examples of real inputs that your system will receive in production. Gather examples of good outputs (from humans or from AI that you've manually verified). Build an evaluation pipeline that runs candidate models against this set and produces metrics you actually care about.

This is expensive. It's also the only approach that works consistently.

Measure What You Care About, Not What's Easy to Measure

Most teams default to measuring accuracy on some held-out set because it's simple to compute. But accuracy often isn't the binding constraint.

Consider: Does it matter more whether the model gets 92% vs. 88% accuracy, or whether the 8-12% of errors are gracefully handled vs. catastrophically wrong? Does it matter more whether the model completes tasks faster, or whether it completes them in a way that reduces downstream human review time?

The metrics that matter are specific to your context. Define them before you start evaluating, not after.

Instrument Production, Not Just Eval

The best evaluation data comes from production systems. Build telemetry that captures not just whether tasks succeeded, but how they succeeded or failed.

When users override AI suggestions, log what the AI suggested and what the user chose instead. When AI-generated code goes through code review, capture the review comments. When AI-completed tasks get escalated to humans, track why.

This feedback loop is more valuable than any benchmark because it's measuring exactly what you care about in exactly the environment where it matters.

Watch What Emerges From This Program

The first outputs from Snorkel's grant recipients won't appear until Q3 2026 at the earliest. But the program structure—quarterly applications, rolling selections, required open-source releases—means there will be a steady stream of new evaluation frameworks and methodologies emerging over the next 18 months.

Set up monitoring for the program's output repository. When frameworks relevant to your use cases appear, evaluate them quickly. The teams building these tools will be actively seeking feedback, and early adopters will have disproportionate influence on how the frameworks evolve.

Where This Leads: The Next 12 Months

Benchmark Fragmentation Before Consolidation

In the near term, expect a proliferation of new benchmarks claiming to measure "real-world performance" or "agentic capability." Most will be poorly designed. Many will be created by labs specifically to make their models look good. The signal-to-noise ratio will get worse before it gets better.

The valuable outcome isn't a single authoritative benchmark. It's the emergence of evaluation methodologies that teams can adapt to their specific contexts. By early 2027, expect to see "evaluation engineering" emerge as a distinct competency within AI teams, separate from model development and MLOps.

Enterprise Buyers Will Demand Better Evidence

The gap between benchmark claims and production reality is increasingly visible to enterprise buyers. CFOs approving seven-figure AI contracts are starting to ask harder questions about what the numbers actually mean.

Vendors who can demonstrate rigorous, production-relevant evaluation will have a significant advantage over those relying on headline benchmarks. Expect RFPs to start including specific requirements around evaluation methodology, not just model performance claims.

The Evaluation Arms Race

As better benchmarks emerge, labs will adapt by optimizing for them. This is inevitable. The question is whether the new benchmarks are harder to game than the old ones.

The dimensions Snorkel's program targets—long-horizon tasks, non-stationary environments, world-modeling—are inherently harder to overfit because they require genuine capability rather than pattern matching. A model can memorize the kinds of answers that score well on a static benchmark. It can't memorize its way through an environment that changes unpredictably.

Open Evaluation as Competitive Advantage

Companies that contribute to open evaluation infrastructure—either through Snorkel's program or independently—will build reputation and influence in the ecosystem. This matters more as the industry matures and trust becomes a differentiator.

The opposite is also true. Labs that resist independent evaluation, or that cherry-pick benchmarks to present their models favorably, will face increasing skepticism from sophisticated buyers.

The Bigger Picture

Snorkel's $3M commitment is modest relative to the scale of AI investment. But the program represents something more significant than its dollar amount suggests.

We're at a pivot point where the AI industry's credibility depends on whether its claims can be independently verified. The benchmark problem isn't just a technical inconvenience—it's an epistemological crisis. When every major lab claims state-of-the-art performance and no one can reliably compare those claims to reality, the entire enterprise of AI evaluation loses meaning.

The market has tolerated this ambiguity because demand for AI capabilities has exceeded supply. Enterprise buyers accepted vendor claims because the alternative was falling behind competitors who were deploying AI regardless of measurement uncertainty. That tolerance is eroding.

Industry coverage of the announcement has focused on the program mechanics—the partners, the grant structure, the focus areas. But the real story is the emergence of institutional infrastructure for accountability in AI evaluation.

The next generation of AI systems will be judged not by their scores on benchmarks designed to be beaten, but by their performance on tasks designed to matter. The teams that prepare for this shift now—by building internal evaluation competency, participating in open evaluation initiatives, and demanding evidence over claims—will be positioned to deploy AI systems that actually work.

The $3M isn't buying better benchmarks—it's buying the foundation for an evaluation ecosystem where benchmark gaming is harder than building genuinely capable systems.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting □